DOCUMENT RESUME

ED 350 318 TM 019 070

AUTHOR Miao, Chang Y.; Kramer, Gene A.
TITLE Detecting Differential Item Functioning Using the Rasch Model with Equivalent-Group Cross-Validation.
PUB DATE Apr 92
NOTE 9p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *College Entrance Examinations; College Students; Comparative Testing; Dental Schools; *Equated Scores; Females; Higher Education; *Item Bias; Item Response Theory; Males; Sampling; *Sex Differences; Test Construction; *Test Items
IDENTIFIERS *Cross Validation; *Rasch Model; T Test

ABSTRACT

An approach to detecting differential item functioning using the Rasch model with equivalent-group cross-validation was investigated. College students taking the Dental Admission Test, were divided by gender (936 females and 1,537 males) into 2 different samples. Rasch analyses were performed on both samples. Data were recalibrated after misfitting persons and items were removed. Resulting difficulties from the two samples were compared, and some potentially biased items were identified. Each sample was then randomly divided into two equivalent samples, resulting in four groups, two of males and two of females. Rasch calibrations were again performed. Items were rank-ordered, and items with the same rankings were selected for equating. Link constants were calculated by selecting one of the four groups as the ground scale. Using the t-test, item-by-item comparisons were conducted within each group pair after the equating. Different difficulty calibrations were compared across the different groups. The significantly different difficulties may have been caused by sampling fluctuation if the t-value was significant when comparing groups with the same gender membership. Finally, identified items were examined for bias. The final set of biased items was then distributed to the test construction committee for possible modifications. One table contains descriptive statistics from the analysis. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED350318

Detecting Differential Item Functioning Using the Rasch Model With Equivalent-group Cross-validation

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

CHANG Y. MIAO

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Chang Y. Miao
American Dental Association

Gene A. Kramer
American Dental Association

A paper presented at the annual meeting of
the American Educational Research Association,
San Francisco, April 1992

TM019070

BEST COPY AVAILABLE

Detecting Differential Item Functioning Using the Rasch Model With Equivalent-group Cross-validation

Abstract

The objective of this study is to investigate an approach to detecting differential item functioning using the Rasch Model with equivalent-group cross-validation. First, the subjects were divided into two different samples according to their gender. Rasch analyses were performed on both samples. After the misfitting persons and items were removed, the data were re-calibrated. The resulting difficulties from the two different samples were compared, and some potentially biased items were identified. After this bias analysis, each sample was randomly divided into two equivalent samples. This resulted in four samples: two groups for female and two for male. The Rasch calibrations were again performed on these four groups. Items were rank-ordered, and the items with the same rankings were selected for equating. Link constants were calculated by selecting one of the four groups as the ground scale. Using the t-test, item by item comparisons were conducted within each group pair after the equating. Different difficulty calibrations were compared across the different groups. The significantly different difficulties may have been caused by sampling fluctuation if the t value was significant when comparing groups with the same gender membership. Finally, identified items were examined for bias. The final set of biased items was then distributed to the test construction committee for possible modifications.

Detecting Differential Item Functioning Using the Rasch Model With Equivalent-group Cross-validation

Objective.

The objective of this study is to investigate an approach to detecting differential item functioning using the Rasch model with equivalent-group cross-validation. It is apparent that detecting biased items is an important part of maintaining an acceptable item bank. There are several methods for detecting biased items that have been discussed in the literature. All the methods recommended have their advantages and disadvantages. For example, the widely used Mantel-Haenszel procedure does not make clear the practical implications of the different alternative hypotheses in test-item studies (Cole & Moss, 1989). Also, IRT methods are complex and require a certain number of subjects in the sample. The Mantel-Haenszel procedure only provides an approximation to the IRT-based methods (Hambleton & Rogers, 1989). Thus, for detecting differential item performance, one of the IRT methods is used in this study. To overcome the complexity of the IRT model, this study is designed to take advantage of the efficiency of detecting biased items using the Rasch model with equivalent-group validation after removing misfitting persons and items.

Theoretical Framework.

The Rasch model has simpler, or perhaps fewer, properties when compared to other IRT methods, but has not been recommended because of problems associated with the confounding of misfit and

bias (Shepard, Camilli, & Willians, 1984). However, under the assumption that the discrimination parameter is a constant, the Rasch model does not involve the problem of explaining difficulty differences calibrated from two different groups when an item is biased. The idea behind the Rasch model is to allow the model to remove the ability distribution of any group used to estimate item difficulties; hence, difficulty estimates should be statistically equivalent for groups distinguished only by their ability distributions (Draba, 1977). Once the data have been "purified" by the appropriate test of fit to the Rasch model, the difficulties estimated from each group for an item should be statistically equivalent except for a single constant of translation that is the same for all items in both tests (Wright, 1977). The separate estimations from two different test sets can be brought together by equating through common items. Using estimated difficulty as the criterion, the common items are defined as those items which have the same difficulty ranked in each group after removing misfit factors. Biased items are investigated after equating; an item is considered biased when it is more difficult when compared to another group. The significant group difference is tested by the 't' statistic (Draba, 1977). Procedurally, this study generates two equivalent groups for different genders to conduct cross-validation analysis (Shepard, 1983). Each analysis is repeated on equivalent samples to confirm the t-statistic result.

Instrument and Methods.

The method is applied to the Quantitative Reasoning Test which contains a total of 50 items. Items consist of computation and word problems. The Quantitative Reasoning Test is a part of the Dental Admission Test battery. The Dental Admission Test is used for graduate admission to dental school programs in the United States. The test is administered to approximately 5,500 college students annually. Four samples were selected from the October 1988 test administration. First, subjects were divided into two different samples according to their gender. Rasch analyses were performed on both samples. The Rasch calibrations were obtained using the BIGSCALE computer program (Wright, Linacre & Schulz, 1990). After the misfitting persons and items were removed, the data were re-calibrated. The criterion for removing the misfitting persons and items was a standardized mean-square outfit statistic greater than 2.5. The resulting difficulties from the two different samples were compared, and some potential biased items were identified.

After this bias analysis, each sample was randomly divided into two equivalent samples. This resulted in four samples: female group 1 (F_1), female group 2 (F_2), male group 1 (M_1) and male group 2 (M_2). The Rasch model analyses were performed again on each of these four groups. Items were rank-ordered, and the items with the same rankings were selected for equating. Link constants were calculated treating group F_1 as the ground scale.

Using the t-test, item by item comparisons were conducted within each group pair. If item estimates differed more than one-half logit for the different calibrations and the t statistic was larger than 2.4, then the item was considered a biased item (Wright & Douglas, 1975; Draba, 1979). Different t values were compared across the different groups. The significantly different difficulties may have been caused by sampling fluctuation if the t value was significant when comparing groups with the same gender membership. Finally, identified items were examined for bias. The final set of biased items was then distributed to the test construction committee for possible modifications.

Results and Discussion.

The first Rasch estimation showed that there are 143 misfitting subjects and six misfitting items. The misfitting items and subjects were excluded. Then, the rest of the sample was divided randomly into equivalent groups. The descriptive statistics are shown in the following table.

Group	Total	Mean logit	Standard Deviation
F_1	474	-0.17	0.78
F_2	463	-0.18	0.78
M_1	764	0.06	0.80
M_2	773	0.04	0.82

The regular practice of bias analysis on female (474+463=937) subjects and male (764+773=1537) subjects found 15 items as biased items. Ten of those 15 items favor males and five favor females. Interestingly, among those ten items favoring males, nine of them are word problems, while two of those items favoring females are word problems. After using equivalent groups in a validating process, the number of biased items were reduced to five items. Four of these five items were identified previously by the regular analysis. Among those five items, four of them favor males. And, among those items favoring males, only one item is dealing with computation. The only item favoring females is a computational item. For the cross-validation approach, ten biased items identified by the first approach showed non-significant t values (less than 2.4). The equivalent group approach removes the confounding sampling error and identifies the biased items more accurately. A closer examination of the content of these five items allows the test construction committee to recognize the possible reasons underlying differential performance. Therefore, the equivalent-group validating process is suggested when the Rasch model is used to judge the fairness of the test items.

References.

Cole, N.S. and Moss, P.A., (1989). Bias in test use. In R. L. Linn (Ed.), Educational Measurement, 3rd ed., 201-220.

Draba, R.E. (1977). The identification and interpretation of

item bias. Education Statistics Laboratory Memorandum 26, University of Chicago, Department of Education.

Hambleton, R.K., Rogers, H.J., (1989). Detecting Potentially biased test items: comparison of IRT Area and Mantel-Haenszel methods. Applied Measurement in Education; v2,n4,313-314.

Shepard, L., Camilli, G., and Williams, D. (1984). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22, 77-105.

Shepard, L., and others (1983). Accounting for statistical artifacts in item bias research. Paper persented at the Annual Meeting of the American Educational Research Association at Montreal, Quebec, 1983.

Wright, B.D., Linacre, J.M., and Schultzm M. (1989). BIGSCALE: A Rasch Model rating scale analysis computer program. Chicago.

Wright, B.D., (1977). Solving measurement problems with the Rasch Model. Journal of Educational Measurement, v14, n2. 97-116.

Wright, B.D., Douglas, G.A. (1975). Best test and self-tailored testing. Education Statistics Laboratory Memorandum 19, University of Chicago, Department of Education.